IML_project

Heikki Nenonen

2022-11-22

Todo

- ~~dummy classifier~~
- class4 -> event/nonevent, week1 exe?
- varianssit mukana/ei mukana? ei one hot -> yksinkertaistaa liikaa ja tarkoitettu kategoriseen dataan
- date? paljon informaatiota, mutta halutaanko muuttujaksi <- opeta 2000-2008, testaa 2009-2011 / kysy slack test_hidden ei - - date, jätetäänkö pois?
- train, test, cv-10?
- itse logisticregression, week2 exe1 <- lasso/ridge
- accuracy, perplexity, week2 exe1
- accuracy of our accuracy? <– malli train+test, vähän parempi kuin pelkkä train?
- class4 -> nonevent/1a/1b/II/
- googlaa mahdollisia malleja

```
# Python with sklearn

import pandas as pd
from sklearn import linear_model

#npf_test = pd.read_csv("initial_data/npf_test_hidden.csv")

npf_train = pd.read_csv("initial_data/npf_train.csv")

npf_train
```

```
##       id        date    class4  ...  UV_B.std   CS.mean    CS.std
## 0      1  2000-01-17        Ib  ...  0.018122  0.000243  0.000035
## 1      2  2000-02-28  nonevent  ...  0.003552  0.003658  0.000940
## 2      3  2000-03-24        Ib  ...  0.272472  0.000591  0.000191
## 3      4  2000-03-30        II  ...  0.451830  0.002493  0.000466
## 4      5  2000-04-04  nonevent  ...  0.291457  0.004715  0.000679
## ..   ...         ...       ...  ...       ...       ...       ...
## 459  460  2011-08-16  nonevent  ...  0.496816  0.002423  0.000425
## 460  461  2011-08-19  nonevent  ...  0.726461  0.002476  0.000902
## 461  462  2011-08-21  nonevent  ...  0.363890  0.003484  0.000457
```

```
## 462  463  2011-08-22  nonevent  ...  0.595032  0.004782  0.001082
## 463  464  2011-08-27  nonevent  ...  0.722553  0.006956  0.000605
##
## [464 rows x 104 columns]
```